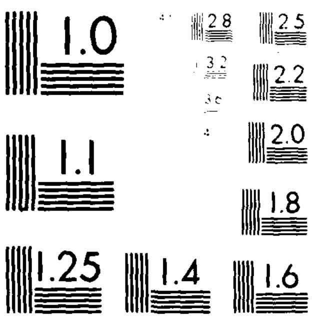
AD-A112 048   MASSACHUSETTS UNIV AMHERST LAB OF PSYCHOMETRIC AND --ETC F/G 5/10
LATENT TRAIT MODEL CONTRIBUTIONS TO CRITERION-REFERENCED TESTIN--ETC(U)
FEB 82 R K HAMBLETON   F33615-79-C-0020
UNCLASSIFIED   AFHRL-TP-81-33   NL

1 OF 1
AD A
112 048

END
DATE
FILMED
04-82
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

AFHRL-TP-81-33

12

AIR FORCE

HUMAN RESOURCES LABORATORY

LATENT TRAIT MODEL CONTRIBUTIONS TO CRITERION-REFERENCED TESTING TECHNOLOGY

By

Ronald K. Hambleton

University of Massachusetts
Laboratory of Psychometric & Evaluation Research
Hills South, Room 152
Amherst, Massachusetts 01003

LOGISTICS AND TECHNICAL TRAINING DIVISION
Technical Training Branch
Lowry Air Force Base, Colorado 80230

February 1982

Final Report

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE,TEXAS 78235

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

ROGER PENNELL
Contract Monitor

ROSS L. MORGAN, Technical Director
Logistics and Technical Training Division

DONALD C. TETMEYER, Colonel, USAF
Chief, Logistics and Technical Training Division

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TP-81-33	2. GOVT ACCESSION NO. AD-A112048	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) LATENT TRAIT MODEL CONTRIBUTIONS TO CRITERION-REFERENCED TESTING TECHNOLOGY		5. TYPE OF REPORT & PERIOD COVERED Final
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Ronald K. Hambleton		8. CONTRACT OR GRANT NUMBER(s) F33615-79-C-0020
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Massachusetts Laboratory of Psychometric & Evaluation Research Hills South, Room 152 Amherst, Massachusetts 01003		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE61102F 2313T209
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE February 1982
		13. NUMBER OF PAGES 30
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Logistics and Technical Training Division Technical Training Branch Air Force Human Resources Laboratory Lowry Air Force Base, Colorado 80230		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

criterion-referenced testing
item characteristic curves
item response theory
latent trait theory
test development

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The goals and results of an 18-month study addressing latent trait model applications to measurement problems arising in criterion-referenced testing are presented. The research studies described in the report cover the following areas: (a) problems with classical test models; introduction to latent trait models, features, assumptions, parameter estimation, and test and item information curves; building tests with latent trait models; (b) latent ability scales—uses, interpretations, and properties; equating test scores for using a common set of norms tables; approaches for addressing the goodness of fit between a latent trait model and a data set; (c) comparing the one-parameter and three-parameter logistic models for ability estimation and decision-making with several test lengths and ability levels; (d) determining the optimal length of criterion-referenced tests with different types of item pools (varying the level of

Item 20 (Continued):

item heterogeneity) and using two different item selection methods; (e) a system to allow instructors to specify required level of measurement precision and obtain information to help them determine test length; (f) comparing the fit of the one-parameter and three-parameter models to 25 sets of test data; and (g) building banks of valid test items.

AFHRL Technical Paper 81-33 February 1982

LATENT TRAIT MODEL CONTRIBUTIONS TO CRITERION-REFERENCED TESTING TECHNOLOGY

By

Ronald K. Hambleton

University of Massachusetts
Laboratory of Psychometric & Evaluation Research
Hills South, Room 152
Amherst, Massachusetts 01003

Reviewed by

James R. Burkett
Chief, Training Systems Section

Submitted for Publication by

Allen J. Partin, LtCol, USAF
Chief, Technical Training Branch

Latent Trait Model Contributions to Criterion-Referenced Testing Technology

Ronald K. Hambleton
University of Massachusetts, Amherst

Two important technologies, criterion-referenced testing and latent trait theory, have emerged in the last 10 years or so and both have the potential for improving Air Force instructional testing. Criterion-referenced testing is the better known of the two, and it is being used widely in the military. Criterion-referenced tests are being used for monitoring student progress in courses, assigning grades, and evaluating courses and training programs. Nevertheless, in spite of the wide-scale use of criterion-referenced tests in the military, many technical problems remain. Two of the most important concerns involve producing technically sound and content-valid test items and determining the number of items for a test.

The second technology, known as item response theory or latent trait theory, has developed more slowly but it is now commonly used by major test publishers, state departments of education, and some test consulting firms in building tests, studying item bias, and equating test scores (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978). Allen Birnbaum, with financial support from the Air Force (see Birnbaum, 1957, 1958a, 1958b), was one of the earliest and most important contributors of statistical theory to the field of latent trait theory.

Latent trait models and related concepts have been used to advance many norm-referenced testing practices and they appear to have the capability for resolving some of the problems associated with criterion-referenced testing. It is surprising to observe then, that latent trait theory has not been used to any considerable extent to address some of the technical matters associated with criterion-referenced testing (see, for example, Hambleton & Cook, 1977).

Purpose

The purpose of this research project was to address several important technological issues and problems associated with criterion-referenced testing via the use of latent trait theory and related concepts. Specifically, seven research reports were prepared to cover the following topics:

1. Problems with classical test models; introduction to latent trait models, features, assumptions, parameter estimation, and test and item information curves; and construction of tests with latent trait models;

2. Latent ability scales — uses, interpretations, and properties; test score equating for using a common set of norms

tables; and approaches for addressing the goodness of fit between a latent trait model and a data set;

3. Comparison of the one- and three-parameter logistic model for ability estimation and decision-making with several test lengths and ability levels;

4. Determination of the optimal length of criterion-referenced tests with several types of item pools (i.e., pools which vary in their level of item heterogeneity) and with two item selection methods;

5. A procedure to allow instructors to specify desired levels of measurement precision and produce tests of necessary lengths;

6. Comparison of the fit of the one-parameter and three-parameter model to 25 sets of test data;

7. Production of banks of valid test items.

The seven reports were prepared to cover the questions outlined in contract F33615-79-C-0020:

a. How can latent trait models be used to estimate examinee domain scores and/or assign examinees to mastery states?

b. What are the "best" methods for equating scores from criterion-referenced tests when there are few test items and a small number of examinees?

c. How should item banks be described and how should test specifications be determined and test items selected for criterion-referenced tests?

d. When learning hierarchies of objectives can be established, how can the information be used with adaptive testing and latent trait models to reduce the amount of necessary testing time?

The fourth question above concerning learning hierarchies was assigned very little attention in the research project. Instead, special attention was devoted to the practical problem of determining appropriate test lengths for criterion-referenced tests. Since the orientation taken in this project was toward solving practical criterion-referenced testing problems for Air Force instructors, as the work progressed, it became clear that an emphasis on the test length determination problem was needed. Determining the number of items is a problem encountered by an instructor everytime he or she constructs a test. While, on the other hand, to date, there has been only limited use of learning hierarchies in instructional testing.

Utilizing learning hierarchies and adaptive testing require the availability of computer terminals for assessment and more items for suitable ability estimation via latent trait models than are usually available.

In a few words, item response theory postulates that (1) underlying examinee performance on a test is a single ability or trait, and (2) the relationship between examinee performance on each item and the ability measured by the test can be described by a monotonically increasing curve. The curve is called an item characteristic curve and it provides the probability of examinees at various ability levels answering the item correctly. In Figure 1 below, two item characteristic curves are shown.

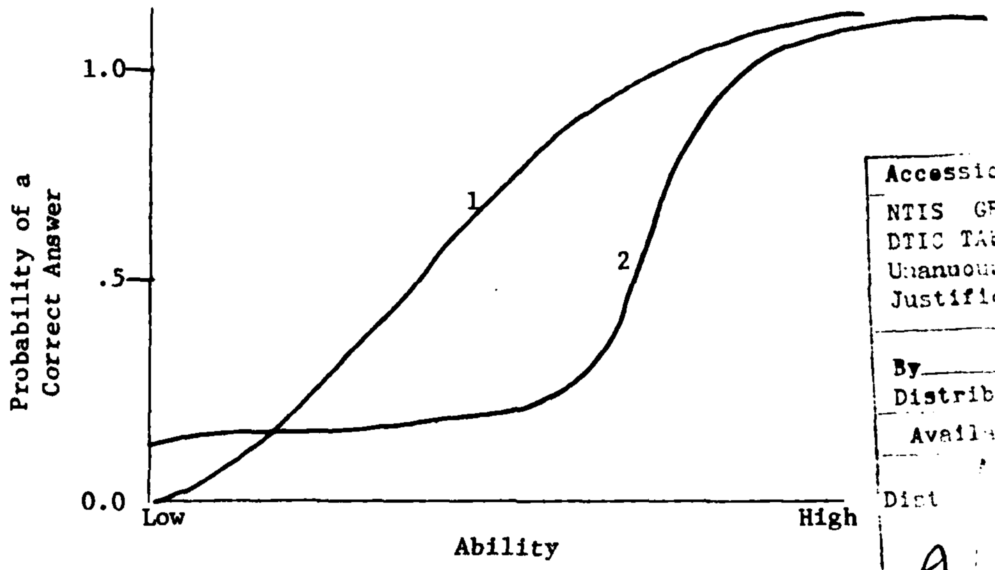


Figure 1. Two item characteristic curves.

It is clear from the figure that the probability of a correct answer depends on the level of examinee ability. Examinees with more ability have higher probabilities for giving correct answers to items than lower ability examinees. Item characteristic curves are typically described by one-, two-, or three-parameter curves. The three item parameters are called item difficulty, item discrimination, and item pseudo-chance level. Items which are shifted to the right end of the ability scale are more difficulty than those shifted to the left end of the ability scale. It is clear from Figure 1 then that item 2 is more difficult than item 1. The slope of an item characteristic curve describes an item's discriminating power. In Figure 1, therefore, item 2 is more discriminating than item 1. Finally, the probability of a very low ability examinee answering an item correctly is the item's pseudo-chance level. With item 1 in the figure, the probability is 0. With item 2 the probability in somewhat higher.

Most item response models, and all of the models which are presently popular, require the assumption that the test items are homogeneous in the sense that they measure a single ability or trait. In addition, it is common to assume that the item characteristic curves are described by one-, two-, or three-parameters, and the corresponding models are referred to as one-, two-, and three-parameter models, respectively. With the three-parameter model, items can vary in their difficulty, discrimination level, and pseudo-chance level. With the two-parameter model, the pseudo-chance level parameter is 0 for all items. With the one-parameter model, not only does the pseudo-chance level parameter have a value of 0 for all items, but all items have a common level of discrimination.

When the assumptions of item response theory can be met in the data sets to which it is applied, at least to a reasonable degree, what is obtained are (1) examinee ability estimates in the pool of items from which the items are drawn that do not depend upon the particular sample of items selected for the test, and (3) item descriptors or statistics (difficulty, discrimination, pseudo-chance level) that do not depend upon the particular sample of examinees from the population of examinees for whom the earlier mentioned item pool is suitable.

What follows in this report are summaries of the seven research papers and the project conclusions.

1. Latent Trait Models and Their Applications to Constructing Tests and Using Test Scores

This initial paper was prepared to provide readers with a non-technical introduction to latent trait models, concepts, and assumptions and to address the important problem of building tests utilizing latent trait models.

What is a latent trait model? It is a model that supposes examinee performance on a test can be predicted (or explained) in terms of one or more characteristics, referred to as traits. A successful latent trait model provides a means of estimating scores for examinees on these traits (Lord & Novick, 1968). Unfortunately, traits cannot be observed directly. They must be estimated (or inferred) from observable examinee performance on a set of test items. This explains the reason for the reference to latent traits or latent abilities. A latent trait model specifies a relationship between the observable test performance of an examinee and the unobservable traits or abilities assumed to underlie performance on a test. The relationship between the "observable" and the "unobservable" quantities is a mathematical function. For this reason, latent trait models are aptly referred to as mathematical models.

There are three primary advantages to latent trait models: (1) assuming the existence of a large pool of items all measuring the same trait, the estimate of an examinee's ability is independent of the particular sample of test items that are administered to the examinee; (2) assuming the existence of a large population of examinees, the descriptors of a test item (for example, item difficulty and discrimination indices) are independent of the particular sample of examinees drawn for the purpose of calibrating the item; and (3) a statistic indicating the precision with which each examinee's ability is estimated is provided. (This statistic is free to vary from one examinee to another.) Of course, the extent to which the three advantages are gained in an application of a latent trait model depends upon the closeness of the "fit" between a set of test data and the model. If the fit is poor, the three desirable features either will not be obtained or obtained in a low degree.

It was determined that there are three important differences between conventional procedures for developing tests and methods that stem from latent trait models: (1) determining item characteristics, (2) selecting items, and (3) estimating test score reliability.

With respect to (1), item analysis techniques involve (a) the characterization of test items and (b) the use of statistical information for revising and/or deleting test items. The major problem with item statistics derived from conventional item analyses is that they are sample dependent. This is unfortunate because these conventional statistics have the advantage of being relatively easy to compute and being estimated relatively precisely with samples of moderate size.

Using conventional item analysis procedures, the detection of "bad" items (for norm-referenced tests at least) is primarily a matter of studying item statistics. A bad item is one that is too easy or too difficult or non-discriminating (i.e., has a low item-total score correlation). Of course, because these statistics are sample dependent, an item may have relatively bad statistics in one sample of students and relatively good statistics in a second group.

The advantages and disadvantages of latent trait models for the purpose of item analysis are almost exactly the reverse of the advantages and disadvantages of conventional item analysis. The item parameters of a latent trait model are sample invariant. But, large sample sizes are required in order to obtain stable estimates of the item parameters and the techniques used to obtain these estimates are complex and difficult to execute, even on a computer.

When applying standard test development techniques to the construction of norm-referenced tests by drawing items from a large pool, items are selected on the basis of two statistics: proportion-correct (difficulty) and the item-total score correlation coefficient (discrimination). Latent trait models provide the test developer not only with sample invariant item parameters but also with a powerful method of item selection. This method involves the use of information curves; i.e., items are selected on the basis of the amount of information they will contribute to the total amount of information supplied by the test.

Finally, when standard test development methods are employed, one or more of the following approaches to the estimation of reliability are used: (1) parallel-form reliability; (2) test-retest reliability; (3) corrected split-half reliability; and (4) internal consistency reliability. All four measures of reliability are sample specific. This unfortunate property of standard estimates of reliability reduces their usefulness. With latent trait models, the analog of reliability and the standard error of measurement is the test information curve. The use of the test information curve as a measure of accuracy of estimation is appealing for at least two reasons: (1) its shape depends only on the items included in the test, and (2) it provides an estimate of the error of measurement at each ability level.

2. Latent Ability Scales, Interpretations and Uses

The purposes of this paper were (1) to discuss the characteristics, interpretations, and uses of ability scales, and (2) to describe several practical methods for assessing the goodness of fit between a test model and a test data set. The latter is important to establish because when the fit between a model and a data set is poor, the expected properties of the ability scale will not exist.

The term "ability" (or latent ability, as it is sometimes called) is a label which is used to designate the trait or characteristic that a test measures. It is broadly defined to include cognitive abilities, achievement variables, basic competencies, personality variables, etc. Rentz and Bashaw (1977) have noted, "The term 'ability' should not be mysterious; it should not be entrusted with any surplus meaning nor should it be regarded as a personal characteristic that is innate, inevitable or immutable. Use of the word 'ability' is merely a convenience."

How is the ability scale obtained? The input to a latent trait analysis, regardless of the choice of model, is the examinee responses to the items in a test of interest. Also, there are a variety of models and ability and item estimation methods that may be chosen. Beyond those important considerations, the central problem becomes one of assigning ability scores to examinees and parameter estimates to items so that there is the maximum agreement possible between the model (as fitted to the data) and the data. This type of analysis is usually carried out with the aid of one of two widely known computer programs, BICAL (Wright & Stone, 1979) or LOGIST (Lord, 1980). In some applications (for example, item banking) estimates of the item parameters are already available. The only problem is to obtain ability estimates. Since the scale on which ability scores are reported is arbitrary, it is common to linearly transform the scores to a convenient scale; for example, in some applications it is convenient to set the mean and standard deviation of the ability scores to 0 and 1, respectively.

One interesting application of latent trait models is to the problem of comparing examinees when they have taken different tests. This problem is considered in detail in the second section of the paper. Why would anyone wish to administer different sets of test items to examinees? One reason is that instructors may wish to administer particular items to examinees because of their diagnostic value. A second reason is that with students who may be expected to do rather poorly or well on a test, better estimates of their abilities can be obtained when test items are selected to match their expected ability levels (Hambleton, 1979).

Latent trait models offer a number of advantages for test score interpretations and reporting but the advantages will be obtained in practice only when there is a close match between the model selected for use and the test data. In a final section of the paper methods for determining goodness of fit were considered.

It was determined that how well a model accounts for a set of test data can be addressed in at least three ways:

a. Determine if the test data satisfy the assumptions of the test model of interest.

b. Determine if the expected advantages derived from the use of a latent trait model (for example, invariant item and ability estimates) are obtained.

c. Determine the closeness of the fit between predictions of observable outcomes (for example, test scores distributions) utilizing model parameter estimates and the test data.

3. Ability Estimation with Three Logistic Test Models

The success of objectives-based instructional programs (i.e., making instructional decisions, and evaluating course effectiveness) depends to a considerable extent upon how effectively criterion-referenced tests are constructed, and how the test scores are used to assess examinee performance levels. While criterion-referenced test development and test score usage have become very popular, because of a shortage of criterion-referenced testing technology, many of the constructed tests do not achieve their full potential. Often test items do not measure the objectives they were developed to measure, too few test items are used in the tests, performance standards are set without due consideration of the relevant issues and consequences, and so on.

Fortunately, the situation has improved. Particularly in the last five years or so there have been a large number of very useful contributions to the criterion-referenced testing literature (Hambleton, 1981). Such contributions have made it possible to develop better criterion-referenced tests and to use the scores in more appropriate ways. For example, much is known about steps for developing criterion-referenced tests, assessing content validity, assembling tests, setting performance standards, and assessing test reliability.

Still, several very important problems remain. For one, what are the best methods for obtaining more accurate estimates of examinees' domain scores (level of performance scores for each objective being tested) and for decreasing the frequency of times examinees are misclassified (assigned to "non-mastery" states when they are "masters" and assigned to "mastery" states when they are "non-masters")? The purpose of this study was to compare the one-, two-, and three-parameter logistic models for estimating domain scores (proportion-correct scores) and making instructional decisions.

The study was conducted using computer simulation methods with tests of several lengths (10, 15, 20, and 40 test items) and several cut-off scores, and the results were compared at different levels of ability (ranging from very low to very high). The steps in the reserach were as follows:

1. Specify the characteristics of a "typical" pool of test items[1] having the following characteristics:

   i. "b" uniformly distributed on the interval [-2.0, +2.0]
   ii. "a" uniformly distributed on the interval [.40, 2.0]
   iii. "c" uniformly distributed on the interval [.15, .25].

2. Select item parameters (b, a, c) from the distribution above for 40 test items.

3. Draw 2,000 examinees from a normal distribution of ability with mean equal to zero and standard deviation equal to one.

4. Simulate the item responses of the 2,000 examinees on the 40 item test. This step produces a 2000 x 40 matrix of item scores.

5. In turn, fit the one-, two-, and three-parameter models using LOGIST to the data set so that item parameters associated with each test model are obtained.

[1] "b" = item difficulty; "a" = item discrimination; "c" = item pseudo-chance level.

6. Next, select one value from each dimension,

   i. ability level (-2.0, -1.5, -1.0, -.5, .0, .5, 1.0, 1.5, 2.0)
   ii. test length (10, 15, 20, 40)
   iii. test model (1, 2, 3)
   iv. cut-off score (.50 above or below the chosen ability level),

   and, then, generate 200 response patterns for the examinee on the test. The response patterns were generated using the "true" item parameters from step 2. Using the estimated item parameters for the model under study, obtain an ability estimate for each response pattern.

   Finally, two statistics were calculated[1]:

$$\frac{\sum_{i=1}^{200} |\theta_i - \theta|}{200} \qquad [1]$$

<u>and</u> Prob $(\theta_i \geq \theta_o \mid \text{if } \theta < \theta_o)$ [2]

or, Prob $(\theta_i < \theta_o \mid \text{if } \theta \geq \theta_o)$ [3]

   Statistic [1] is the average absolute deviation of estimates ($\hat{\theta}$) about the true ability (θ). Statistic [2] is the proportion of times that ability estimates exceeded the cut-off score when the examinee was a non-master ($\theta<\theta_o$). This is known as the false-positive error rate. Statistic [3] is the proportion of times the ability estimates for an examinee were below the cut-off score when the examinee was a master ($\theta\geq\theta_o$). This is known as the false-negative error rate.

7. All possible combinations of 6i, ii, iii, and iv were studied and comparisons among the three logistic models were carried out.

The results of the study were clear:

1. Not surprisingly, better domain score estimates and smaller false-positive and negative error rates were observed when the tests were lengthened. The results also showed that the two-parameter model was somewhat better than the one-parameter model for improving the quality of estimates and decisions with the shortest tests.

[1] θ = true ability; $\hat{\theta}_i$ = ability estimate for examinee i; θ_o = cut-off score.

In conclusion, it was noted in the paper that many reasons have been offered for not using the two- and three-parameter models —

i. they require too much computer time.

ii. the computer program in common use (LOGIST) places strict constraints on the "a" and "c" parameters to obtain convergence.

iii. The c's are poorly estimated.

But, the study by Hutten (1981) clearly shows that the computer costs associated with LOGIST are not unreasonable. Also, while constraints are placed on the parameter estimates, the item and ability estimates arrived at with the computer program (LOGIST) are close to the true parameter values (Lord, 1980). And, it has been shown that when sufficient numbers of low ability examinees are present, the c parameters can be properly estimated. What this study does show, however, is that when making descriptions or decisions on the basis of criterion-referenced test performance, all three models give similar results except at the low end of the ability continuum where the three-parameter model does a substantially better job. The more general models do function a little better overall but not enough to justify their use in most military classroom settings for assessing examinee ability. It is, of course, important to stress that this conclusion should not be generalized to other possible applications of latent trait models.

4. Determining the Optimal Lengths of Criterion-Referenced Tests

If criterion-referenced testing is to achieve its full potential, in addition to other desirable technical characteristics, criterion-referenced test scores must lead to decisions that are "consistent." That is, a high percentage of examinees must be classified into the same mastery state with a parallel-form (or a readministration) of the test, or the resulting decisions are of limited value. Unfortunately, there are no practical methods to assist the test developer in determining the number of items required to achieve some desired level of decision-consistency. Existing methods are either based on unreasonable assumptions, are highly conservative, or fail to consider important factors. For example, the well-known generalized Spearman-Brown formula is useful for determining the desirable length of a norm-referenced test but it is of limited value in building criterion-referenced tests. The Spearman-Brown formula uses the correlation between scores on parallel-forms of a test whereas with criterion-referenced tests, interest is centered on the consistency of decision-making across parallel-form (retest) administrations of a test. The "correlation of scores" and "consistency of decisions resulting from the use of scores" will, in general, have different values.

Three specific solutions to the test length determination problem have been offered in the criterion-referenced testing literature. All of the solutions have shortcomings. One solution requires the use of the simple binomial test model and the approximate true score for the examinee to be tested (Millman, 1973). But, it is seldom the case that all items from a domain of content will have similar or identical item difficulty levels for an examinee (one assumption of the simple binomial test model), and since the purpose of testing is to assess examinee true score, there will be at least some occasions when a suitable true score estimate cannot be found. In a second solution, test length is determined so as to insure the probabilities of correctly classifying non-masters and masters exceed some desired levels (Wilcox, 1976). This solution has substantial merit but it does lead to the selection of highly conservative (i.e., longer) test lengths. That is, considerably longer tests are used than are needed with many examinees to achieve the desired probabilities of correct classifications. Eignor and Hambleton (1979) offered a third solution which seems to have some merit as well as several shortcomings. Test developers must specify an expected true score distribution (this is easier to do than guessing a particular examinee's true score) and their solution incorporates the use of the compound-binomial test model (a somewhat more plausible test model than the model used in the Millman solution). Also, in the Eignor-Hambleton solution, test lengths are chosen to achieve some desired level of decision consistency for the specified true score distribution and so test lengths are shorter than those obtained with the Wilcox method. But, the Eignor-Hambleton solution has several serious shortcomings. For one, the heterogeneity of the item pool used in test development is not considered. It should be considered because heterogeneity of the item pool will have a direct impact on the required test length. Longer tests will be needed when the item pool is heterogeneous.[1] Also, Eignor and Hambleton do not consider the importance of the approach chosen for building parallel-forms. Generally, longer tests are required when parallel-forms are constructed by randomly sampling items from the pool than by careful matching of item statistics in the parallel-forms.

In view of the several shortcomings of existing methods for determining test lengths and the importance of a satisfactory solution to the problem for a suitable criterion-referenced testing system, a study was designed and conducted to further investigate the test length determination problem.

[1] The situation is analogous to the problem of (say) estimating the percent of persons in a population of interest who support a particular legislative action. More persons will need to be sampled to obtain a valid estimate of the preference in the population when members of the population are highly diverse than when members of the population are very similar. In a diverse or heterogeneous population, estimates are apt to vary substantially from one sample to the next, especially when the chosen samples are small. With a homogeneous population, the estimates will vary less from one sample to the next.

The principal purpose of the study was to address, via the use of computer simulation techniques and latent trait models, the relationships among: (1) test length, (2) decision consistency, (3) kappa (another useful measure of decision consistency that takes into account the amount of agreement due to chance), (4) the heterogeneity of an item pool, and (5) the item selection method used to construct parallel-forms. In addition, tables were produced to aid the test developer in selecting optimal test lengths.

The principal analyses for the study were carried out with three simulated item pools constructed with the aid of latent trait models which varied in terms of their range of item difficulty indices. In the first pool, the range of p values was 0[1]; in the second pool, the range was moderate (about .50); and in the third pool, the range was high (about .80). These three item pools were chosen to represent the extremes of those pools that might be found in practice as well as one that could be described as "typical." Test length was varied between two and 20 items. These limits are reasonable since it is seldom that more than 20 items are used to measure a single objective in a criterion-referenced test.

Items were drawn from the pools for the purpose of building parallel-forms of a test by using two methods: in the "randomly-parallel method," items were drawn at random from the item pool to create parallel-forms. In the "strictly-parallel method," corresponding items in each form were chosen to be statistically equivalent. Both item selection methods *(to a substantial degree) are presently* in common use in the building of criterion-referenced tests. An additional variable manipulated in the study was the amount of examinee guessing behavior: two values were considered, .00 and .20. The .20 value implies that examinees with low levels of ability had about a 20% chance of guessing the answers to items correctly, while the .00 value implies minimal guessing.

Examinee item performance on the parallel-form of each test was simulated via the use of the three-parameter logistic test model and a random number generator. Once item scores were available, test scores were calculated and mastery assignments made by a comparison of examinee test scores to the minimum standard of performance. With each item pool, method of item selection, level of guessing, and test length, examinees were assigned to mastery states based upon their performance on the parallel-forms of the test being simulated. Next, for a group of 300 simulated examinees, decision consistency and kappa were calculated. Three-hundred examinees were simulated in each run to insure that stable estimates of decision consistency and kappa were obtained. In total, 120 computer runs were made (3 item pools x 2 methods of item selection x 2 levels of guessing x 10 test lengths).

[1] Another way to say that the range of p values was 0 is to say that all of the test items were equally difficult.

The cut-off score (or minimum standard of performance as it is sometimes called) used to assign examinees to mastery states was situated near the center of the domain score distribution. This cut-off score value had the advantage of producing lower bound estimates of decision consistency for the group of examinees being tested. In practice, therefore, it would be expected that estimates of decision consistency would be at least as high and probably higher than the values reported in the results section of the paper.

The results reported in the study can be summarized as follows:

1. When an item pool is homogeneous, considerably shorter tests are needed to achieve desired levels of decision consistency.

2. There is considerable advantage to building parallel-forms via matching items statistically. This is especially true when the tests of interest are on the short side.

3. The average level of the discriminating power of items in a pool has a substantial influence on the required lengths of tests (i.e., shorter tests suffice when test items are highly discriminating).

4. When guessing is a substantial factor in test performance, decision validity is lowered.

5. The many tables and figures in the paper provide a basis for test developers to determine the number of items they need in particular testing situations.

The importance of this paper is that it (1) provides useful tables for test developers to assist them in the difficult technical problem of determining the optimal number of test items; and (2) shows the impact of item pool characteristics, item selection methods, and examinee guessing behavior on test lengths to achieve desired levels of decision consistency and kappa. To date, these three factors have received almost no attention in the criterion-referenced testing literature. This is unfortunate because of the impact of these three factors on the test length determination problem. In view of the need for a satisfactory solution to the test length determination problem, methods and results described in this paper should be of considerable value to criterion-referenced test developers.

5. A Method for Determining the Length of Criterion-Referenced Tests Using Reliability and Validity Indices

Criterion-referenced tests are used to determine an examinee's status with respect to well-defined domains of behavior (Hambleton & Eignor, 1979; Popham, 1978). Construction of a criterion-referenced test usually involves (among other things) drawing a representative sample of items from a pool of items which measures the content domain of interest. Of central importance in the test development process is the determination of the number of items to be included. The length of the test (or subtests if several objectives are measured in a test) is directly related to the usefulness of the scores. In general, short tests lead to less reliable and valid scores than longer tests. Longer tests, however, while generally resulting in more precise estimates of ability, require more testing time and may cause examinee fatigue if they become very long. Also, since it is often the case that several objectives are assessed in a single criterion-referenced test, practical considerations argue against a large number of items per objective. It is important, therefore, that criterion-referenced tests contain enough items to yield scores with desired levels of reliability and validity without requiring excessive amounts of testing time.

The primary use of criterion-referenced tests is to assign examinees to categories or states reflecting levels of performance in relation to the objectives measured in a test. When mastery decisions are being made, it is possible to determine test length in relation to the number of misclassification errors which can be tolerated. The purpose of this paper was to describe a system, implemented with the aid of a computer, which can be used to determine test lengths which would lead to specified levels of classification errors. In the paper, several procedures for determining test lengths were reviewed and the computer-assisted system for determining test length was presented.

The FORTRAN program, TESTLEN, was designed to allow users to specify local conditions and to simulate test performance. By simulating several possible test lengths and cut-off scores, users can obtain estimates of various statistics of interest. The values obtained may then be used to make decisions regarding optimal test lengths. As a result, requirements for test development or item selection are clarified.

TESTLEN can be used to simulate parallel administrations of several criterion-referenced tests. Characteristics of the tests (test length, cut-off score, distribution of item parameters) and characteristics of the examinee pool (number of examinees, distribution of domain scores) are under user control. Also, under user control is the number of replications of each parallel form administration to be simulated. Multiple replications allow users to determine the stability of the results.

Output from the program includes the following information for each replication:

1. item difficulties (p-values)

2. item-test (subtest) correlations

3. item b, a, and c values (latent trait simulations only)

4. the number of examinees

5. decision consistency

6. kappa (or Cohen's kappa)

7. decision accuracy

8. chance agreement

9. proportion of examinees in each mastery classification.

In each situation the mean, range, and standard deviation of decision consistency, kappa, and decision accuracy across the replications are reported.

Figure 2 contains a portion of an output from the program. The simulation utilized the three-parameter logistic latent trait model to generate the responses of 100 examinees to randomly parallel forms of a 10-item test. The 100 examinees were distributed normally on the latent ability scale with mean 0.0 and standard deviation 1.0. Item difficulties were specified to range from -2.00 to +2.00; item discrimination values ranged from +0.40 to +2.00; pseudochance level values ranged from +0.15 to +0.25. The cut-off score was set at 0.00 (the center of the ability distribution) and the advancement score was 5 items correct.

It is not always possible to accurately specify characteristics of examinees and item pools. In such cases test developers will probably want to err on the side of conservatism since it may be better to have a few extra items than to err on the short side and have an unacceptable number of classification errors. The following recommendations were offered in the paper to provide guidelines for producing conservative test lengths. First, use sample sizes similar to the number of examinees to be tested. Larger samples will yield more stable estimates of reliability and validity, but test developers need to know the expected range of these statistics in their situation. Second, when in doubt about the distribution of domain scores, and when a conservative solution is desired (i.e., a solution which will insure a desirable level of test reliability and validity even if an unexpected test score distribution is obtained) it is best to center the distribution close to the cut-off score. The closer the distribution is to the cut-off score, the more classification errors will result. Thus, more items will be required to reach acceptable levels of decision accuracy. Third, if characteristics of the item pool are not established, specify heterogeneous pools. This selection will lead to more conservative estimates of test length.

TESTLEN simulates parallel-form administrations of criterion-referenced tests. Some options of the program allow the user to choose

SIMULATED ITEM-TEST CORRELATIONS—FORM A
REPLICATION NUMBER 2

	Conventional			Latent Trait	
Item	Item	Item-Test			Pseudo-
Number	Difficulty	Correlation	Difficulty	Discrimination	Chance
1	.43	.5739	1.723	.614	.216
2	.42	.6677	1.080	1.484	.245
3	.74	.6601	-1.551	1.821	.234
4	.67	.7513	-.832	1.969	.169
5	.34	.5182	1.795	1.567	.235
6	.65	.7748	-.836	1.272	.219
7	.41	.5349	1.211	.921	.168
8	.69	.6407	-.842	.821	.239
9	.36	.3786	1.379	.625	.200
10	.69	.7287	-.942	1.294	.182
Mean	.54	.6229	.218	1.239	.211
Range	.40	.3962	3.346	1.355	.077

SIMULATED ITEM-TEST CORRELATIONS—FORM B
REPLICATION NUMBER 2

	Conventional			Latent Trait	
Item	Item	Item-Test			Pseudo-
Number	Difficulty	Correlation	Difficulty	Discrimination	Chance
1	.46	.7246	1.003	1.242	.152
2	.45	.5987	1.016	.650	.191
3	.50	.6961	.752	1.259	.191
4	.49	.6402	1.315	1.829	.213
5	.44	.6406	1.594	.494	.164
6	.73	.5224	-.857	.474	.200
7	.54	.6271	.624	1.507	.223
8	.48	.5525	1.290	1.216	.239
9	.67	.6680	-.792	1.503	.221
10	.51	.6753	.727	1.860	.193
Mean	.53	.6345	.667	1.203	.199
Range	.29	.2023	2.451	1.386	.087

WHEN DIFFICULTY IS 0.00 or 1.00, ITEM-TEST CORRELATIONS CANNOT BE COMPUTED. THESE ITEMS ARE DENOTED BY CORRELATIONS OF 9.9999.

SUMMARY STATISTICS FOR REPLICATION NUMBER 2

# of	# of	Cut-Off	Advance	Dec-		Dec-		Mas-	Mas-	Non-	Non-
Examinees	Items	Score	Score	Con	Kappa	Acc	Chance	Mas	Non	Mas	Non
100	10	0.00	5	.86	.72	.91	.50	.49	.11	.03	.37

```
                    SUMMARY OF 2 REPLICATIONS

                          Dec-Con              Kappa              Dec-Acc
  # of       # of                   Std.                 Std.                 Std.
Examinees   Items   Mean   Range    Dev.   Mean  Range   Dev.   Mean  Range   Dev.

   100        10     .82    .09     .064   .63    .17    .122    .89   .05    .035
```

Figure 2. Sample output from TESTLEN. This is the second of two replications in which latent trait theory was used to simulate the performance of 100 examinees on randomly parallel forms of a 10-item test.

between the construction of randomly or statistically parallel tests. If two tests are to be developed by randomly selecting items from an item pool, the user should specify randomly parallel tests. If, however, the tests are to be matched on item statistics, the user should choose the alternate option.

Most of the options included in TESTLEN rely on a random number generator. Users will possibly have to modify the program to conform to the random number generator at their facility.

Finally, the test length determination problem must be solved for each objective (or competency) on the test for which mastery decisions will be made.

6. Fitting the One- and Three-Parameter Models to a Variety of Tests

This study was undertaken to highlight some similarities and differences between the one- and three-parameter logistic models. Specifically, the purpose of the study was to compare the Rasch (or one-parameter) logistic model with the Birnbaum (or three-parameter) logistic model by fitting the models to empirical data.

Proponents for both models have asserted that their model is most appropriate for describing test behavior. This research was designed to provide Air Force personnel with information to help in selecting latent trait models for estimating ability. The study examined fit of the models to empirical data. Model fit was systematically analyzed in terms of deviations from latent trait model assumptions present in the data. This comparison is important because of the potential ramifications, legal or otherwise, that could result when the assumptions of the models have not been met. The study also examined the appropriateness of the models in situations where test lengths were very short or where few examinees were available for estimating item parameters. This part of the study also provided information on the precision of ability estimates based on short tests. Finally, comparative costs for estimating parameters by the two models were reported. While cost should not be the primary reason for selecting one model over the other, expenses are an important issue today because of shrinking budgets. Because the research study was exploratory in nature, no specific hypotheses were tested, rather, the study sought to provide information in the following areas:

1. Which methods can be used to determine that empirical data do not violate the underlying assumptions of latent trait models? The assumptions include unidimensionality, equality of item discriminations (one-parameter model), and minimal guessing (one-parameter model). Information in this area was obtained from a substantial review of the literature. Various procedures were explored on a trial basis, and those selected were critically analyzed. Recommendations were made for how model assumptions can be tested.

2. How is goodness-of-fit defined and what statistical, graphical, and practical procedures can be employed to determine model fit? Three measures of fit were used in the study. Outcomes based on each measure were compared and suggestions offered for future research.

3. Do latent trait models fit tests developed by conventional methods? Which model demonstrates better fit to empirical data? Fit statistics and graphic evidence are presented for the fit of the one- and three-parameter models to 25 empirical data sets. Results based on various methods of fit were compared.

4. How do deviations from latent trait model assumptions affect fit of data to the latent trait models? Are the models robust to violations in their assumptions? For both models, fit was explored in terms of unidimensionality. For the one-parameter model, fit statistics were examined when equality of item discriminations and minimal guessing assumptions were violated in the data.

5. How well do latent trait models fit data when estimates of ability are made on short tests? Three measures of precision for short tests were used: Pearson correlations, Spearman rank order correlations, and average absolute differences (AAD).

6. How well do latent trait models fit data when item parameter estimates are based on very small examinee samples? Pearson correlations, Spearman correlations, and AAD statistics were used to explore precision of item parameters from small samples.

7. What are the comparative costs (in terms of computer time and expense) for obtaining parameter estimates of the two latent trait models? CPU time and cost were tallied and compared for parameter estimation under each model.

Item response data for 25 tests were obtained from a variety of sources to make comparisons of the fit of the one- and three-parameter models to empirical data. Data were from multiple-choice tests designed for the measurement of achievement or aptitude. The tests covered a broad range of contents, formats, difficulty levels, and examinee sample characteristics. Five of the tests were used as a subset to explore the effects of sample size and test length on precision of latent trait parameter estimates, while all 25 tests were used to explore questions of model fit.

Samples of 1,000 examinees were drawn from each data set. An item analysis and a factor analysis were performed with each test. Conventional item statistics were used to roughly approximate the degree to which each data set deviated from the latent trait model assumptions. For example, the standard deviation of item point-biserial correlations provided a rough indication of the extent to which sets of test items differed in their discriminating power. Then, item and ability parameters were estimated with each model. These estimated parameters were substituted for true parameters to make predictions about number-correct score distributions from each model. Predicted distributions were compared with observed raw score distributions by statistical and graphical procedures. Measures of fit were correlated with indices of violation of model assumptions to examine model robustness. Then, precision of parameter estimates from small samples and from short tests were explored with correlational techniques. Finally, computer times and costs for parameter estimation by the two models were compared.

The results of the study demonstrated that latent trait theory is at least adequate for describing outcomes on cognitive tests. Both aptitude and achievement measures constructed with conventional testing methodologies displayed good fit to both of the latent trait models.

The one-parameter model compared favorably with the three-parameter model in this study. On 50 percent of the tests, the one-parameter model fit data as well as the three-parameter model. At least for the purpose of estimating ability, the one-parameter model proved to be almost as advantageous as the three-parameter model.

The lack of unidimensionality in data was found to have severe consequences on model fit. As data tended towards multidimensionality, fit for both the one- and three-parameter models was significantly impaired. The results suggest that unidimensional subsets of items be formed prior to application of latent trait models. Factor analysis was suggested as a method for assessing dimensionality.

In this study, the one-parameter model fit was not affected by the presence of heterogeneous item discrimination indices. The outcomes of this study demonstrated that highly acceptable estimates of ability and item difficulty can be obtained with the one-parameter model from tests with only 20 items and from samples as small as 250 examinees. Although good ability estimates can be obtained with the three-parameter model when tests have only 20 items, pseudo-chance level item parameter estimates and item discrimination estimates from samples of 250 examinees were not adequate. The results of this study supported Lord's (1980) suggestion that 1,000 examinees are required for obtaining good estimates of item discrimination. It would seem that even bigger samples are needed to estimate the guessing parameters properly or alternately, a sample should be selected with a disproportionately large number of examinees at the lower end of the ability scale.

It was also shown that when ability scores were estimated with "known" item parameters, computer expenses for the one- and three-parameter models were about the same. When item parameters were also estimated, the cost of estimation with the three-parameter model was considerably more than with the one-parameter model but both costs were not high ($70.00 maximum for 40 items and 1,000 examinees). With the exception of an initial estimation of item parameters, latent trait models are customarily used to obtain ability estimates. The difference between costs for the two models in the long run, therefore, would seem to be negligible.

7. Building a Bank of Valid Test Items

Item banks consist of substantial numbers of items that are matched to course objectives and can be used by instructors to build tests on an "as needed" basis. When a bank consists of content valid and technically sound items, the instructor's task of building tests is considerably easier and the quality of tests is usually higher than when an instructor prepares his/her own test items. Item banks, especially those for which a scale has been developed to statistically describe test items, offer considerable potential to the Air Force:

- Instructors can easily build tests to measure objectives of interest.
- Instructors, within the limits of an item bank, can produce tests with the desired number of test items per objective.
- If item banks consist of content-valid and technically sound items, test quality will usually be better than instructors could produce if they were to prepare the test items themselves.

It seems clear that in the future item banks will become increasingly important to the Air Force because of the potential they hold for improving the quality of classroom tests and reducing the time spent by instructors in building their tests.

The purpose of this part of our research program was to provide instructors with guidelines for preparing objectives and test items that are consistent with the emerging technology associated with criterion-referenced testing. While the central thrust of the present effort concerned the investigation of latent trait models, without valid test items, any use of latent trait models would be limited in value. Therefore, this paper was prepared to provide Air Force instructors with guidelines for preparing their objectives and producing test items to measure their objectives.

The paper is divided into five sections. Item banks are defined first along with their uses. Methods for specifying content domains, or writing objectives, are addressed in section two. The generation of test items is considered in section three. In section four, a system for conducting technical and content reviews of test items is offered. A conclusion is provided in section five.

An item bank is an organized group of test items that can be retrieved to form a carefully planned test with prespecified characteristics. Conceptually, item banking is a simple process. It begins by collecting test items that have been created by item writers to explicitly assess some clearly defined set of skills or competencies. Next, these items are stored usually by a computer along with their item statistics. Other information is routinely held in the system revealing the extent to which items have been used in the past, and further, an updating process constantly takes place enabling the item bank user to receive up-to-date psychometric data.

Several methods for preparing objectives are reviewed in the paper with "domain specifications" being the one most highly recommended. Many researchers in the field of measurement today advocate the use of domain specifications for defining course content (Berk, 1980; Popham, 1975, 1980; Hambleton, 1981). A domain specification is an expanded statement of an objective which provides information regarding content, types of responses and distractors, and other information deemed necessary for clarification and specification (Popham, 1978). Typically, there are five parts of a domain specification: (1) General Description, (2) Sample Item, (3) Stimulus Attributes, (4) Response Attributes, and (5) Specification Supplement.

Once the course content has been selected and specified, the test items can be generated. During this writing process close attention must be given to the domain specifications to assure that the test items do "tap" behaviors in the set of behaviors explicated by the DSs (Hambleton, Swaminathan, Algina, Coulson, 1978; Hambleton, 1981). The item writer must generate items that assess the objectives they are intended to measure and at the same time, strive to develop test items which are technically sound. Guidelines for choosing appropriate item formats and rules for writing technically sound items were offered in the paper.

No matter how carefully item writers attempt to follow the standard item writing guidelines, test items still must be reviewed for technical quality before they are stored in a bank for later use by instructors. The review process is not a simple one. It takes time to review each test item and make appropriate revisions. It is essential that a review procedure follow the initial item writing process.

Also we cannot assume that the items written from domain specifications measure the behaviors they were designed to assess. A second review process then becomes necessary. That is, although item writers have attempted to generate items that reflect the content circumscribed by domain specifications, methods must be employed to assure that item-objective congruence exists. There are two principal approaches that have been advocated in the past. They include judgmental techniques and empirical methods. Empirical methods were not discussed in the paper. Attention was focused on the use of judgment for checking item-objective congruence.

In summary, the usefulness of an item bank is directly dependent upon the quality of the test items. Unless the test items stocked in a bank represent the content domains they were designed to assess, and are of good technical quality, they are of limited value to instructors. In the paper practical steps for generating suitable test items were offered. The steps are summarized in Figure 3. It is expected that by following the steps the Air Force will be able to stock their item banks with carefully designed test items.

Conclusions

The seven papers described in this report cover a wide range of topics, issues, and technical developments. Their contributions can be summarized as follows:

- Instructional material introducing latent trait models, concepts, and assumptions (Report 1) and ability scales (Report 2).
- Instructional material concerning the preparation and review of objectives and test items for item banks (Report 7).
- Some guidelines for building tests using latent trait models (Report 1).
- A computer simulation solution incorporating latent trait models for the problem of determining criterion-referenced test lengths. The solution includes complete directions for its use and a Fortran IV computer program to carry out the required analyses (Report 4 and 5).
- Procedures for addressing the goodness of fit between a latent trait model and a test data set (Report 2 and 6).
- Steps for equating tests for the purpose of using a common set of norms tables (Report 2).
- A comparison of the three logistic test models with respect to the two principal uses of criterion-referenced test scores: descriptions and decisions (Report 3).
- Extensive comparisons of the fit of the one- and three-parameter logistic models to a variety of data sets (Report 6).
- Cost figures on using the logistic test models (Report 6).

In sum, the results from this study indicate that latent trait models will be useful to instructors who wish to do a better job of building criterion-referenced tests and using the scores. Also, at the classroom level where examinee sample sizes will be small (even after accumulating item response data over a year or two) and small gains in measurement precision will be of limited value, the results in this study indicate that the one-parameter logistic model will suffice. The small gains in measurement precision which, in theory, are obtained when the three-parameter model rather than the one-parameter model is

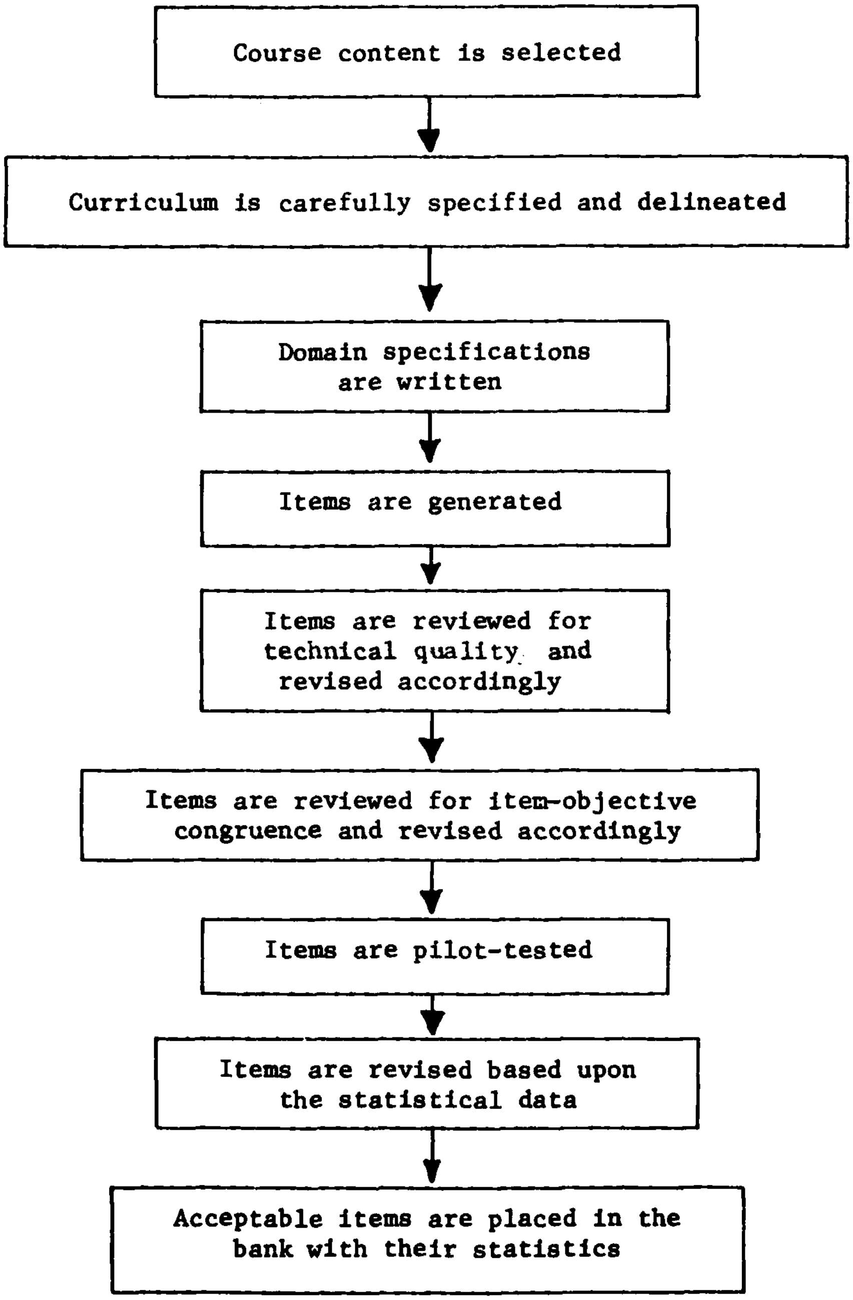

Course content is selected
Curriculum is carefully specified and delineated
Domain specifications are written
Items are generated
Items are reviewed for technical quality and revised accordingly
Items are reviewed for item-objective congruence and revised accordingly
Items are pilot-tested
Items are revised based upon the statistical data
Acceptable items are placed in the bank with their statistics

Figure 3. Steps for producing a bank of test items.

used, do not seem sufficiently important at the classroom level to justify the use of the more complex model. And, even the small gains obtained in this study from the use of the three-parameter model are unlikely to be obtained in practice because the sample sizes available to instructors are (typically) too small to permit proper estimation of the item statistics. It is important, however, that unwarranted generalizations not be made from the results of this study. When examinee sample sizes are larger than those considered in this study, or when even small gains in measurement precision are desired, or when problems other than ability estimation are of central importance, it is entirely possible that the three-parameter model will prove to be more useful than the one-parameter model.

References

Berk, R. A. (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Birnbaum, A. Efficient design and use of tests of a mental ability for various decision-making problems. Series Report No. 58-16. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1957.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

Birnbaum, A. Further considerations of efficiency in tests of a mental ability. Technical Report No. 17. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

Eignor, D. R., & Hambleton, R. K. Effects of test length and advancement score on several criterion-referenced test reliability and validity indices. Laboratory of Psychometric and Evaluative Research Report No. 86. Amherst, MA: School of Education, University of Massachusetts, 1979.

Hambleton, R. K. Latent trait models and their applications. In R. Traub (Ed.), Methodological developments: New directions for testing and measurement (No. 4). San Francisco: Jossey-Bass, 1979.

Hambleton, R. K. Advances in criterion-referenced testing technology. In C. Reynolds and T. Gutkin (Eds.), Handbook of school psychology. New York: Wiley, 1981.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1979.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hutten, L. Fitting the one- and three-parameter models to a variety of tests. Laboratory of Psychometric and Evaluative Research Report No. 116. Amherst, MA: School of Education, University of Massachusetts, 1981.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Popham, W. J. Educational evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1975.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Popham, W. J. Domain specification strategies. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Rentz, R. R., & Bashaw, W. L. The National Reference Scale for Reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Wilcox, R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

Wright, B. D., & Stone, M. H. Best test design. Chicago: MESA Press, 1979.

DATE
ILMED

4-8